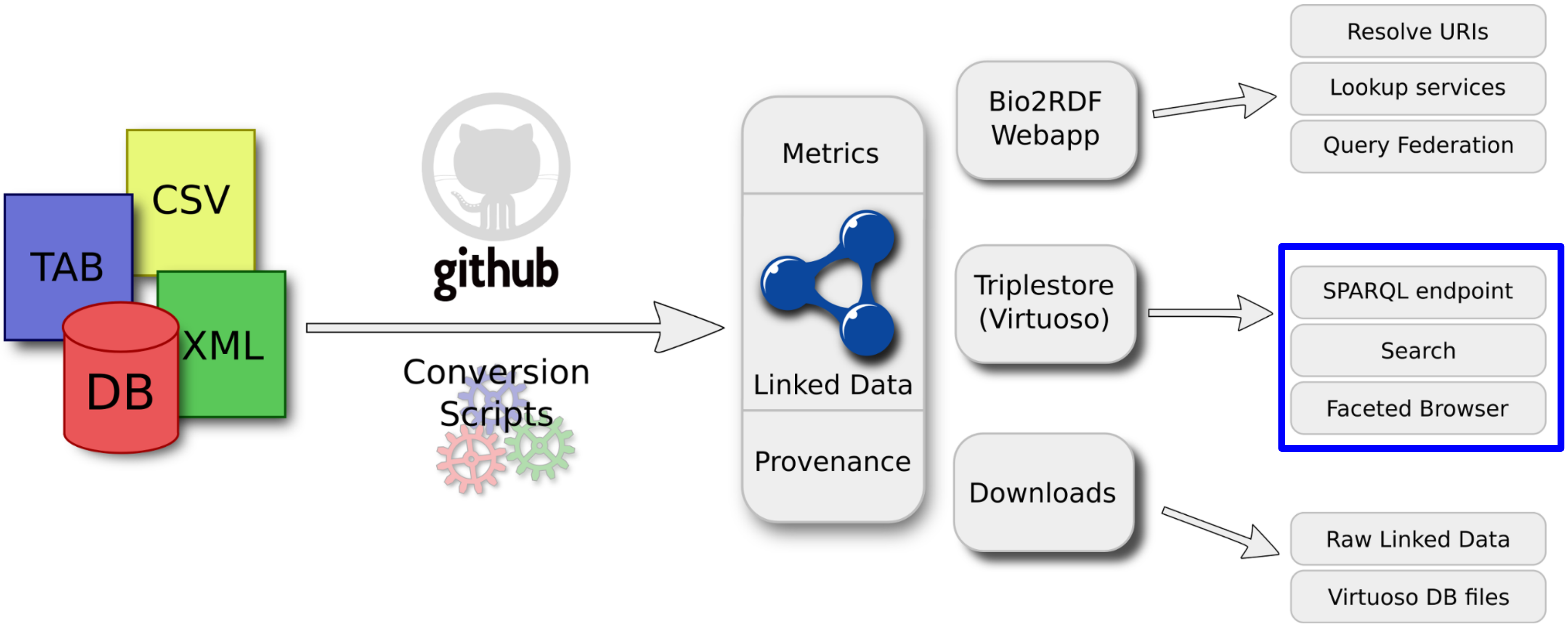
Querying Bio2RDF data

Tutorial @ ICBO 2013

Tutorial Roadmap



SPARQL: The query language of the Semantic Web

- **SPARQL**: **S**PARQL **P**rotocol **A**nd **Q**uery **L**anguage
- SPARQL (“sparkle”) is a W3C recommendation that is part of the semantic web stack
- A SPARQL query allows you to search linked data based on the structure of the triples it contains
- SPARQL can be used to explore the structure of RDF graphs and to transform linked data

Anatomy of a SPARQL query

- SPARQL queries have a regular structure composed of the following parts:
    - Prefix declarations: Shortcuts for URIs used in the query (*e.g.* rdf, rdfs, bio2rdf)
    - Dataset definition: RDF graph to query (support for this option is SPARQL endpoint engine dependent)
    - Result clause: Data returned by the query
    - Query pattern: Graph pattern used to search the RDF data
    - Query modifiers: Limiting, ordering, other forms of result rearrangements

Anatomy of a SPARQL query

```
#comments can be included
PREFIX prefixA: <http://example.org/prefixA#>
PREFIX prefixB: <http://example.org/prefixB:>
SELECT ...
FROM <http://example.org/myDataset>
WHERE {
    ...
} LIMIT 10
```

Federated SPARQL queries over >1 endpoint use the SERVICE keyword

```
PREFIX prefixA: <http://example.org/prefixA#>
PREFIX prefixB: <http://example.org/prefixB:>
SELECT ...
FROM <http://example.org/myDataset>
WHERE {
    SERVICE <http://somewhere.org/sparql> {
        ...
    }
} LIMIT 10
```

Four SPARQL query variants

SELECT: SQL style result set retrieval. Lets you specify the variables you wish to retrieve from the data.

CONSTRUCT: Create a custom RDF graph based on a query criteria. Triples can be extracted verbatim as they exist in the queried triple store or re-constructed to create new RDF data.

ASK: Tests whether the triplestore or graph contains the specified statement. Returns TRUE or FALSE.

DESCRIBE: Returns all of the triples that contain a specified resource.

EXAMPLE: SELECT

Data from Bio2RDF Gene dataset:

```
<http://bio2rdf.org/geneid:19> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://bio2rdf.org/geneid_vocabulary:Gene> .
<http://bio2rdf.org/geneid:19> <http://bio2rdf.org/geneid_vocabulary:has_symbol> "ABCA1" .
<http://bio2rdf.org/geneid:19> <http://bio2rdf.org/geneid_vocabulary:has_description> "ATP-binding cassette, sub-family A (ABC1), member 1" .
<http://bio2rdf.org/geneid:19> <http://bio2rdf.org/geneid_vocabulary:has_taxid> <http://bio2rdf.org/taxon:9606> .
```

Query: Get taxonomic identifier and description for a specific gene symbol

```
PREFIX gene_vocab: <http://bio2rdf.org/geneid_vocabulary:>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?gene ?geneDescription ?taxid
WHERE {
        ?gene gene_vocab:has_symbol "ABCA1" .
        ?gene gene_vocab:has_description ?geneDescription .
        ?gene gene_vocab:has_taxid ?taxid .
}
```

EXAMPLE: CONSTRUCT

Data from Bio2RDF Gene dataset:

```
<http://bio2rdf.org/geneid:19> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://bio2rdf.org/geneid_vocabulary:Gene> .

<http://bio2rdf.org/geneid:19> <http://bio2rdf.org/geneid_vocabulary:has_symbol> "ABCA1" .

<http://bio2rdf.org/geneid:19> <http://bio2rdf.org/geneid_vocabulary:has_description> "ATP-binding cassette, sub-family A (ABC1), member 1" .

<http://bio2rdf.org/geneid:19> <http://bio2rdf.org/geneid_vocabulary:has_taxid> <http://bio2rdf.org/taxon:9606> .
```

Query: Construct dc:identifier triple for an NCBI gene from description

```
PREFIX dc:http://purl.org/dc/terms/
PREFIX gene_vocab: <http://bio2rdf.org/geneid_vocabulary:>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
CONSTRUCT {
        ?gene dc:description ?geneDescription .
} WHERE {
        ?gene rdf:type gene_vocabulary:Gene .
        ?gene gene_vocab:has_symbol "ABCA1" .
        ?gene gene_vocab:has_description ?geneDescription .
}
```

EXAMPLE: ASK

Data from Bio2RDF DrugBank dataset:

```
<http://bio2rdf.org/drugbank_resource:DB00072_DB00563> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>     <http://bio2rdf.org/drugbank_vocabulary:Drug-Drug-Interaction .

<http://bio2rdf.org/drugbank_resource:DB00072_DB00563> <http://www.w3.org/2000/01/rdf-schema#label>  "DDI between Trastuzumab and Methotrexate - Trastuzumab may increase the risk of neutropenia and anemia. Monitor closely for signs and symptoms of adverse events. [drugbank_resource:DB00072_DB00563]" .

<http://bio2rdf.org/drugbank:DB00072> <http://bio2rdf.org/drugbank_vocabulary:is-ddi-interactor-in> <http://bio2rdf.org/drugbank_resource:DB00072_DB00563> .

<http://bio2rdf.org/drugbank:DB00563> <http://bio2rdf.org/drugbank_vocabulary:is-ddi-interactor-in> <http://bio2rdf.org/drugbank_resource:DB00072_DB00563> .
```

Query: Is there a drug-drug interaction between trastuzumab and methotrexate?

```
PREFIX drugbank_vocab: <http://bio2rdf.org/drugbank_vocabulary:>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
ASK WHERE {
        ?ddi rdf:type drugbank_vocab:Drug-Drug-Interaction .
        <http://bio2rdf.org/drugbank:DB00072> drugbank_vocab:is-ddi-interactor-in ?ddi .
        <http://bio2rdf.org/drugbank:DB00563> drugbank_vocab:is-ddi-interactor-in ?ddi .
}
```

EXAMPLE: DESCRIBE

Data from Bio2RDF PharmGKB dataset:

```
<http://bio2rdf.org/pharmgkb:PA443997> rdf:type <http://bio2rdf.org/pharmgkb_vocabulary:Disease> .
<http://bio2rdf.org/pharmgkb:PA443997> rdfs:label "Ehlers-Danlos Syndrome [pharmgkb:PA443997]" .
<http://bio2rdf.org/pharmgkb:PA443997> rdfs:seeAlso <http://bio2rdf.org/mesh:0004535> .
<http://bio2rdf.org/pharmgkb:PA443997> rdfs:seeAlso <http://bio2rdf.org/umls:C0013720> .
<http://bio2rdf.org/pharmgkb:PA443997> rdfs:seeAlso <http://bio2rdf.org/snomed:3A398114001> .
<http://bio2rdf.org/pharmgkb:PA443997> owl:sameAs <http://bio2rdf.org/pharmgkb:00072f176862ae5012d717f2858fcf03> .
<http://bio2rdf.org/pharmgkb:PA443997> <http://bio2rdf.org/pharmgkb_vocabulary:name> "Ehlers-Danlos Syndrome" .
<http://bio2rdf.org/pharmgkb:PA443997> <http://bio2rdf.org/pharmgkb_vocabulary:synonym> "Cutis Elastica" .
<http://bio2rdf.org/pharmgkb:PA443997> <http://bio2rdf.org/pharmgkb_vocabulary:synonym> "Cutis elastica" .
<http://bio2rdf.org/pharmgkb:PA443997> <http://bio2rdf.org/pharmgkb_vocabulary:synonym> "Cutis hyperelastica" .
<http://bio2rdf.org/pharmgkb:PA443997> <http://bio2rdf.org/pharmgkb_vocabulary:synonym> "Danlos disease" .
<http://bio2rdf.org/pharmgkb:PA443997> <http://bio2rdf.org/pharmgkb_vocabulary:synonym> "Cutis hyperelastica
dermatorrhexis " .
<http://bio2rdf.org/pharmgkb:PA443997> void:inDataset <http://bio2rdf.org/bio2rdf_dataset:bio2rdf-pharmgkb-20121015> .
```

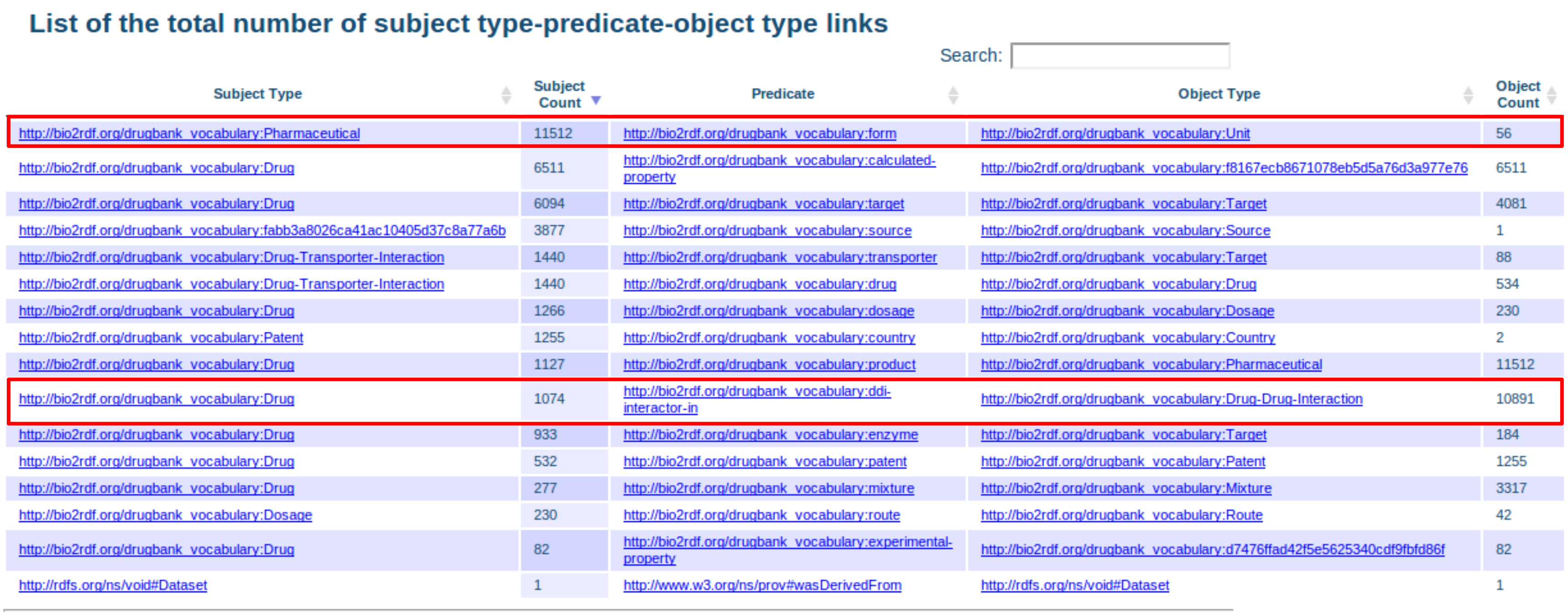
Query: Get all triples involving the PharmGKB resource for Ehlers-Danlos Syndrome

```
DESCRIBE <http://bio2rdf.org/pharmgkb:PA443997>
```

Bio2RDF summary metrics can be used to develop SPARQL queries

- Each Bio2RDF endpoint contains summary metrics about the dataset, including:
  - unique predicate-object links and their frequencies
  - unique predicate-literal links and their frequencies
  - unique subject type-predicate-object type links and their frequencies
  - unique subject type-predicate-literal links and their frequencies
- These can inform SPARQL query development by describing the links that exist between entities of a given type

Bio2RDF summary metrics can be used to develop SPARQL queries



List of the total number of subject type-predicate-object type links

Search:

Subject Type	Subject Count	Predicate	Object Type	Object Count
http://bio2rdf.org/drugbank_vocabulary:Pharmaceutical	11512	http://bio2rdf.org/drugbank_vocabulary:form	http://bio2rdf.org/drugbank_vocabulary:Unit	56
http://bio2rdf.org/drugbank_vocabulary:Drug	6511	http://bio2rdf.org/drugbank_vocabulary:calculated-property	http://bio2rdf.org/drugbank_vocabulary:f8167ecb8671078eb5d5a76d3a977e76	6511
http://bio2rdf.org/drugbank_vocabulary:Drug	6094	http://bio2rdf.org/drugbank_vocabulary:target	http://bio2rdf.org/drugbank_vocabulary:Target	4081
http://bio2rdf.org/drugbank_vocabulary:fabb3a8026ca41ac10405d37c8a77a6b	3877	http://bio2rdf.org/drugbank_vocabulary:source	http://bio2rdf.org/drugbank_vocabulary:Source	1
http://bio2rdf.org/drugbank_vocabulary:Drug-Transporter-Interaction	1440	http://bio2rdf.org/drugbank_vocabulary:transporter	http://bio2rdf.org/drugbank_vocabulary:Target	88
http://bio2rdf.org/drugbank_vocabulary:Drug-Transporter-Interaction	1440	http://bio2rdf.org/drugbank_vocabulary:drug	http://bio2rdf.org/drugbank_vocabulary:Drug	534
http://bio2rdf.org/drugbank_vocabulary:Drug	1266	http://bio2rdf.org/drugbank_vocabulary:dosage	http://bio2rdf.org/drugbank_vocabulary:Dosage	230
http://bio2rdf.org/drugbank_vocabulary:Patent	1255	http://bio2rdf.org/drugbank_vocabulary:country	http://bio2rdf.org/drugbank_vocabulary:Country	2
http://bio2rdf.org/drugbank_vocabulary:Drug	1127	http://bio2rdf.org/drugbank_vocabulary:product	http://bio2rdf.org/drugbank_vocabulary:Pharmaceutical	11512
http://bio2rdf.org/drugbank_vocabulary:Drug	1074	http://bio2rdf.org/drugbank_vocabulary:ddi-interactor-in	http://bio2rdf.org/drugbank_vocabulary:Drug-Drug-Interaction	10891
http://bio2rdf.org/drugbank_vocabulary:Drug	933	http://bio2rdf.org/drugbank_vocabulary:enzyme	http://bio2rdf.org/drugbank_vocabulary:Target	184
http://bio2rdf.org/drugbank_vocabulary:Drug	532	http://bio2rdf.org/drugbank_vocabulary:patent	http://bio2rdf.org/drugbank_vocabulary:Patent	1255
http://bio2rdf.org/drugbank_vocabulary:Drug	277	http://bio2rdf.org/drugbank_vocabulary:mixture	http://bio2rdf.org/drugbank_vocabulary:Mixture	3317
http://bio2rdf.org/drugbank_vocabulary:Dosage	230	http://bio2rdf.org/drugbank_vocabulary:route	http://bio2rdf.org/drugbank_vocabulary:Route	42
http://bio2rdf.org/drugbank_vocabulary:Drug	82	http://bio2rdf.org/drugbank_vocabulary:experimental-property	http://bio2rdf.org/drugbank_vocabulary:d7476ffad42f5e5625340cdf9fbfd86f	82
http://rdfs.org/ns/void#Dataset	1	http://www.w3.org/ns/prov#wasDerivedFrom	http://rdfs.org/ns/void#Dataset	1

http://download.bio2rdf.org/release/2/drugbank/drugbank.html

Bio2RDF summary metrics can be used to develop SPARQL queries

```
PREFIX drugbank_vocabulary: <http://bio2rdf.org/drugbank_vocabulary:>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?ddi ?d1name
WHERE {
   ?ddi a drugbank_vocabulary:Drug-Drug-Interaction .
   ?d1 drugbank_vocabulary:ddi-interactor-in ?ddi .
   ?d1 rdfs:label ?d1name .
   ?d2 drugbank_vocabulary:ddi-interactor-in ?ddi .
   ?d2 rdfs:label ?d2name .
   FILTER (?d1 != ?d2)
}
```

Results: http://bit.ly/14qGfUh

Example Bio2RDF SPARQL queries

Bio2RDF query: Retrieve diseases associated with the BRCA1 gene

```
PREFIX ctd_vocab: <http://bio2rdf.org/ctd_vocabulary:>
SELECT ?disease ?diseaseLabel
FROM <http://bio2rdf.org/ctd>
WHERE {
   ?assoc rdf:type ctd_vocab:Gene-Disease-Association .
   ?assoc ctd_vocab:gene <http://bio2rdf.org/geneid:672> .
   ?assoc ctd_vocab:disease ?disease .
   ?disease rdfs:label ?diseaseLabel .
}
```

Results: http://bit.ly/162NM9L

Bio2RDF federated query: Retrieve GO function labels from BioPortal for a gene in NCBI gene

```
SELECT *
WHERE {
     <http://bio2rdf.org/geneid:3253304> <http://bio2rdf.
org/geneid_vocabulary:function> ?goFunction .
     SERVICE <http://bioportal.bio2rdf.org/sparql> {
          ?goFunction rdfs:label ?label .
    }
}
```

Results: http://bit.ly/13D20SR

Bio2RDF query: Count all the biochemical reactions in the BioModels database involved in "protein catabolic process"

```
SELECT ?go ?label count(distinct ?x)
WHERE {
   {
    # get all the biochemical reactions specifically labelled with protein catabolic
process
    ?go rdfs:label ?label .
    FILTER regex(?label, "^protein catabolic process")
    service <http://biomodels.bio2rdf.org/sparql> {
     ?x <http://bio2rdf.org/biopax_vocabulary:identical-to> ?go .
     ?x a <http://www.biopax.org/release/biopax-level3.owl#BiochemicalReaction> .
    }
   } UNION {
    # get all the biochemical reactions that are more specific than "protein catabolic
process"
    ?go rdfs:label ?label .
    ?go rdfs:subClassOf ?tgo OPTION (TRANSITIVE) . # get all the subclasses of the
target to term
    ?tgo rdfs:label ?tlabel .
    FILTER regex(?tlabel, "^protein catabolic process")
    service <http://biomodels.bio2rdf.org/sparql> {
     ?x <http://bio2rdf.org/biopax_vocabulary:identical-to> ?go .
     ?x a <http://www.biopax.org/release/biopax-level3.owl#BiochemicalReaction> .
    }
  }
}
```

Results: http://bit.ly/14qGWwC

Use the VOS Faceted Browser to explore Bio2RDF data

- Explore types and attributes
- Free text search

Explore Bio2RDF data on your own!

http://download.bio2rdf.org/release/2/release.html